FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1
Organización de Datos

[7506] Organización de Datos
Cátedra Argerich
Primer cuatrimestre de 2021

Alumno:	PAUSELLI, Fabrizio Dante
Número de padrón:	103418
Email:	fpauselli@fi.uba.ar
Alumno:	Fernandez Marchitelli, Camila
Número de padrón:	102515
Email:	cfernandezm@fi.uba.ar
Alumno:	SAN MARTÍN, Nicolás
Número de padrón:	104320
Email:	nsanmartin@fi.uba.ar
Repositorio:	OrganizacionDeDatos-Tp1

Índice

1. Introducción

En el año 2015 Nepal fue afectado por el terremoto Gorkha, un sismo que registró una magnitud de 7.8 en la escala Richter y tuvo su epicentro en la ciudad de Kathmandu. Aproximadamente 600,000 estructuras en el centro y pueblos aledaños fueron dañadas o destruidas. Un análisis posterior al sismo llevado por la Comisión Nacional de Planeamiento de Nepal comunicó que la pérdida total económica ocasionada por el terremoto fue de aproximadamente $7 mil millones (USD; NPC, 2015).

El presente trabajo practico tiene como objetivo realizar un Análisis Exploratorio de Datos, viendo si existe alguna relación entre las edificaciones y los daños que sufrieron.

Teniendo en cuenta que Nepal es un país de 30 millones de habitantes en el cual, aproximadamente, el 45 % de su población es pobre y se encuentra en el puesto 108 del ranking de PIB mundial, consideraremos este factor en determinados análisis.

2. Hipótesis

Listaremos los elementos que creemos que tienen mas influencia en los daños que sufrieron las edificaciones, previo a realizar el análisis.

- Ubicación con respecto al epicentro, cuanto mas cerca, mayor daño.
- Edad de la edificación, cuanto mas años tenga, sufrirá mayor daño.
- Altura, a mayor, menores daños, tal vez por que están mejores preparados para este tipo de eventualidades como un terremoto.
- Material de construcción, si es un material bueno, sufrirá menos daño, y sera el que predominara en las construcciones con menor daño.
- Tipo de superficie, cuanto mejor sea la condición, menor daño sufrirá.

3. Variables a estudiar

Se decidió estudiar las relaciones entre variables, y las que nos parecieron mas significativas fueron 'count_families', 'damage_grade', 'plan_configuration' y las relacionadas con el tipo de material de construcción ('has_superstructure_{material}').

4. Supuestos

Los supuestos que consideramos al realizar el trabajo fueron los siguientes:

1. Las edificaciones que tienen una antigüedad de 995 años, por ser las únicas que presentan un valor de tal magnitud, nos hace pensar que representa un numero por defecto para aquellas que no se pudo obtener la edad. Serán analizadas por separado al final del trabajo.

5. Aclaraciones

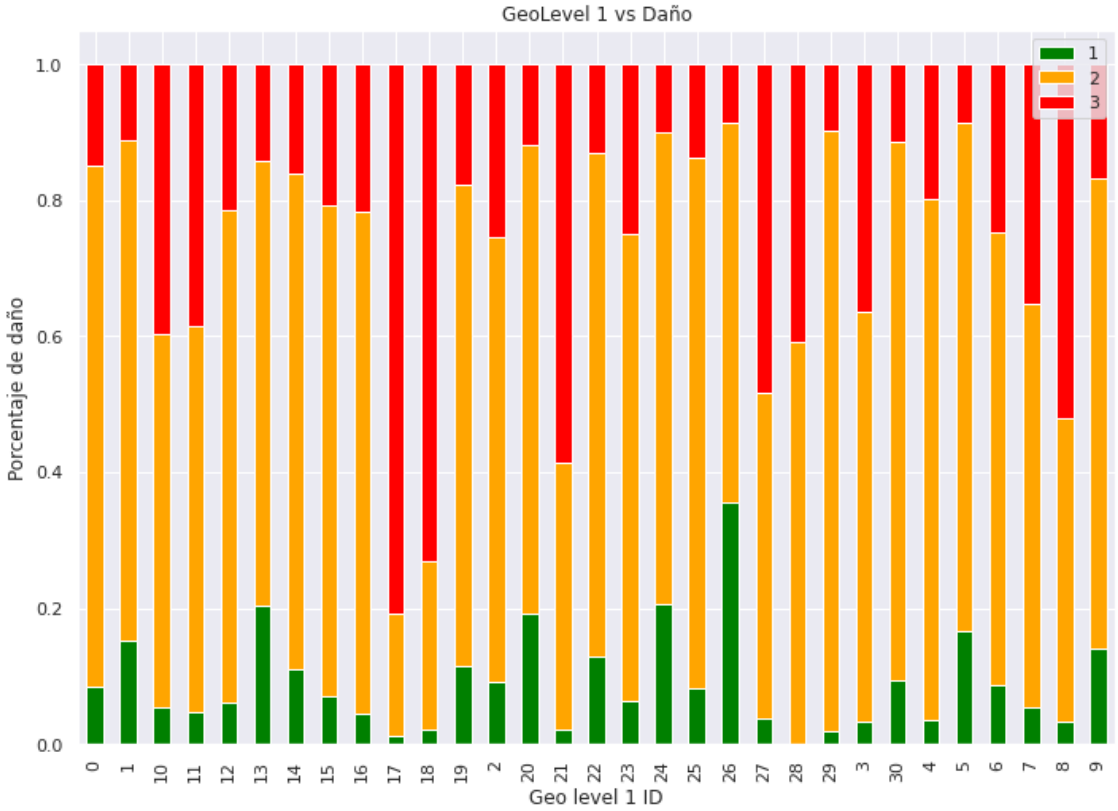
Aclaración de notación:

- Nivel de daño 1 = Poco daño
- Nivel de daño 2 = Daño intermedio
- Nivel de daño 3 = Mucho daño

6. Empieza el análisis

6.1. Primer vistazo de zonas más afectadas

En el gráfico que se muestra a continuación podemos observar el porcentaje del tipo de daño que sufrieron aquellas edificaciones que pertenecen a la zona de Geo Level 1.



En principio se puede notar que aquellas zonas que tuvieron las peores consecuencias fueron la 17 (donde hay 21813 edificaciones, que representa un 8.3 % del total) cuyo daño de tipo 3

predomina por amplia diferencia al daño de tipo 2 y 1, al igual que la 18 (representa el 1.2 %) la 8 y 27 (más del 50 % fue de tipo 3, contando con el 8 % y 5 % del total de edificios). Otra que también nos llamó la atención es la ID 29 (representa un %1 aproximadamente, que tiene mas del 80 % de daño 2)

En las demás regiones las consecuencias fueron mayoritariamente de daño 2. No se observa ninguna región donde el daño de tipo 1 supere al 2 o 3, pero la 26 es una de las que menor daño 3 sufrió y la que más porcentaje de daño 1 tuvo en relación a las otras zonas.

Esta distribución de los daños nos hace pensar que las zonas mencionadas anteriormente, junto a otras no nombradas que también sufrieron bastante daño 2 y 3, se encontraban cerca del epicentro del terremoto o eran zonas con construcciones muy frágiles.

6.2. Cantidad de Familias

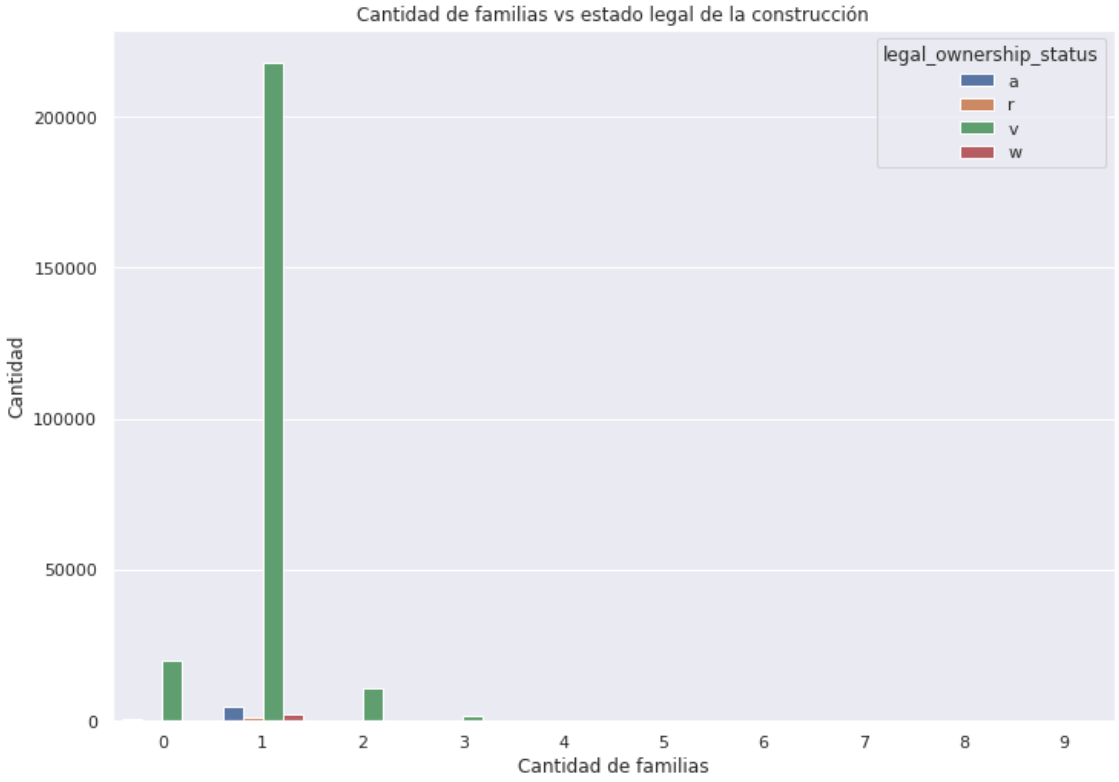
En este apartado estudiaremos las relaciones que tienen las siguientes variables con la cantidad de familias:

- legal_ownership_status
- damage_grade
- geo_level_1_id
- geo_level_2_id
- geo_level_3_id
- plan_configuration

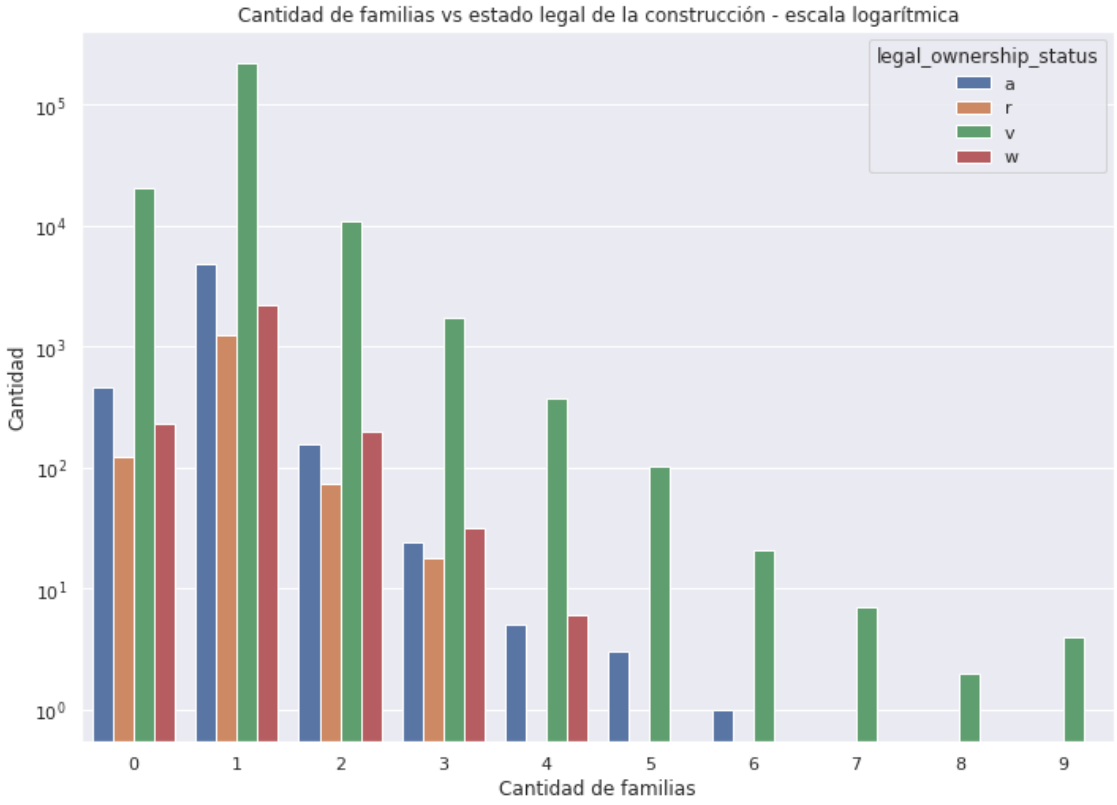
Las preguntas que nos hicimos con respecto a estas variable, en general no estuvieron relacionadas de manera directa al daño, pero nos permitió analizar otras cosas, como cuáles son las zonas con mas familias, o una relación indirecta del nivel socioeconómico de las mismas al analizar el estado legal de las construcciones utilizadas para uso familiar.

6.2.1. ¿Hay alguna relación entre el estado legal de la propiedad y el número de familias?

Cuando quisimos ver el estado legal de la construcción con respecto a la cantidad de familias que viven en la misma, vimos que, con varios órdenes de magnitud, lo que más se veía eran grupos monofamiliares viviendo en construcciones de tipo "v".



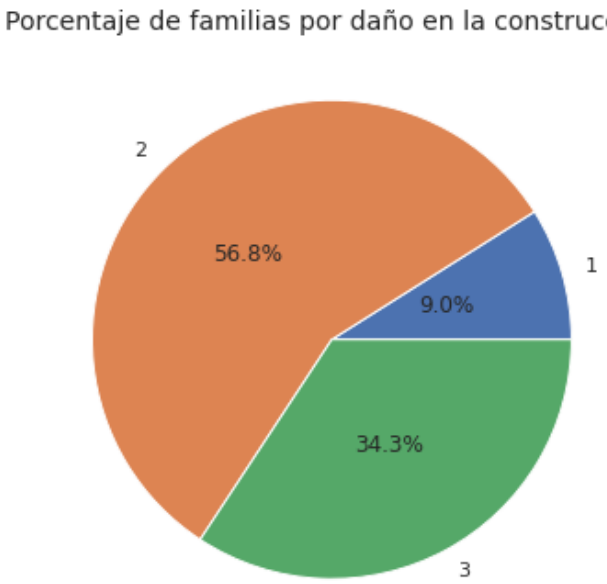
Para poder entender mejor estos datos, decidimos graficarlo en escala logarítmica, donde de todas formas se ve que las construcciones de tipo 'v' predominan para cualquier cantidad de familias, seguido por las de tipo 'a' y 'w'.



6.2.2. ¿Cuántas familias perdieron completa o parcialmente su vivienda?

Aunque no agregue información relevante para hacer predicciones, de todas formas nos pareció interesante analizar la proporción de familias dado el daño que recibió el edificio donde residían.

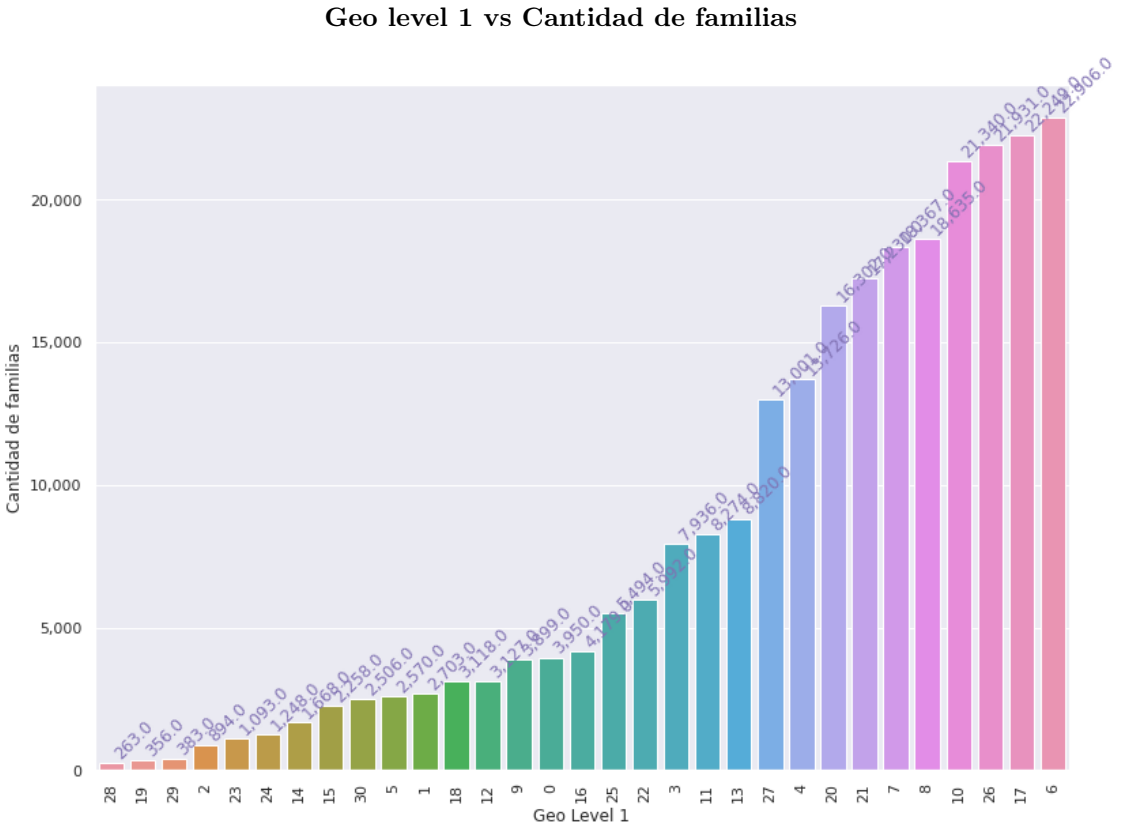
Lo que se puede ver, es que sólo el 9 % vivía en construcciones que recibieron daño leve, siendo los que sufrieron daño moderado la mayor cantidad, con mas de 56 %



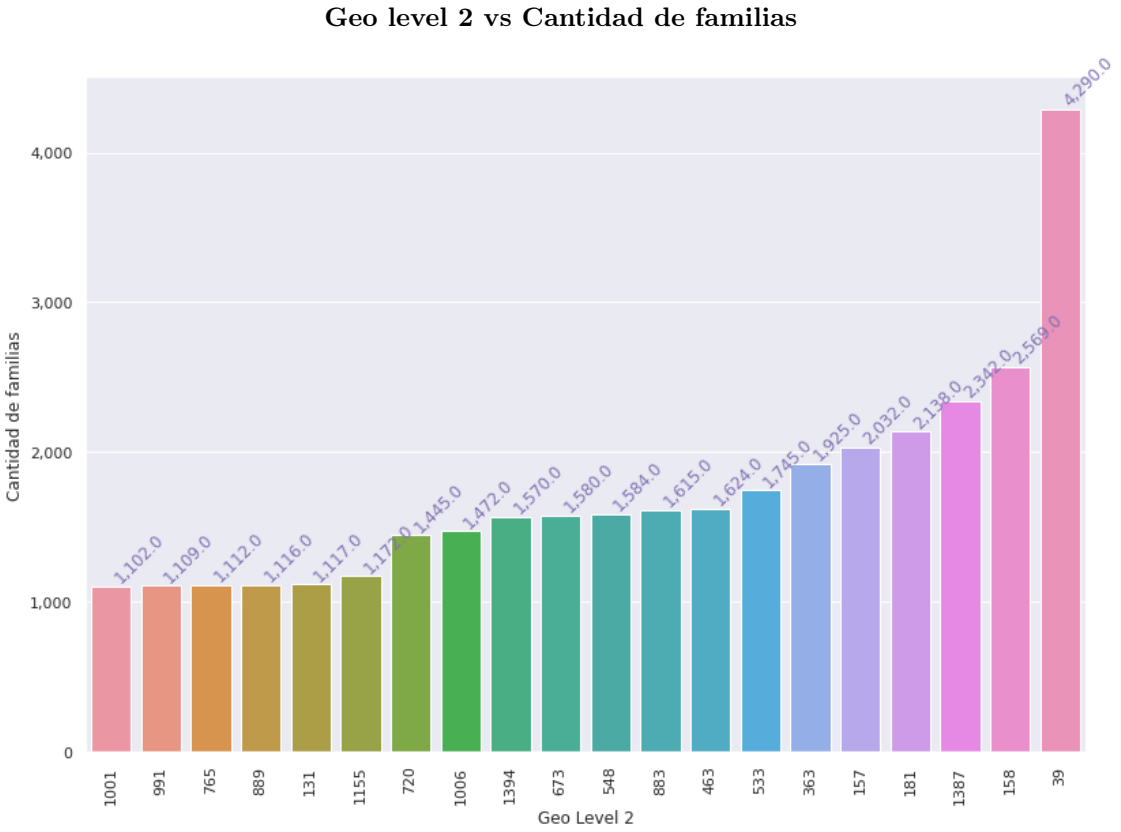
6.2.3. ¿Cómo se relaciona el número de familias con la región geográfica?

En este momento, decidimos analizar la cantidad de familias viviendo en cada región geográfica, y ver cuáles serían los 'geo level' con mas proporción de familias (esto no necesariamente significa los lugares más habitados, porque podrían ser muchas familias de pocos miembros).

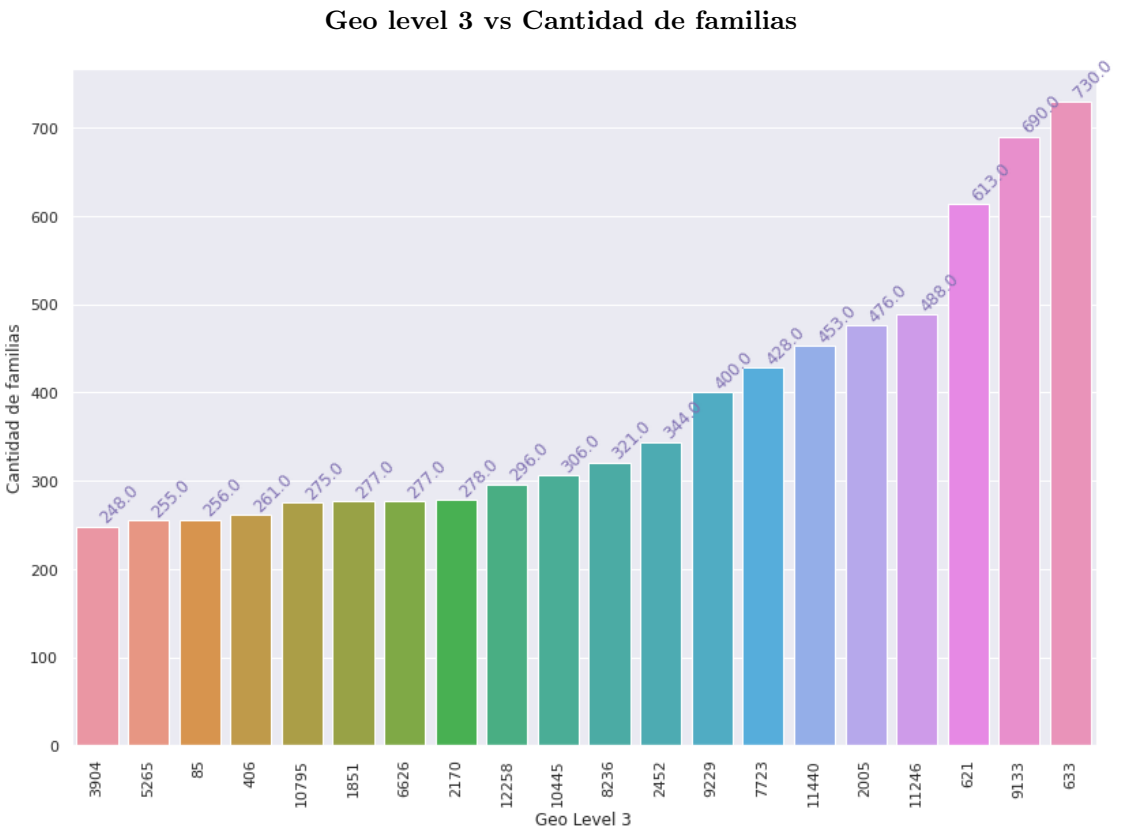
Para 'Geo level 1', se ve que la región con más familias fue la 6



Para 'Geo level 2', se ve que la región con más familias fue con bastante diferencia, la 39



Para 'Geo level 3', se ve que la región con más familias fue la 633



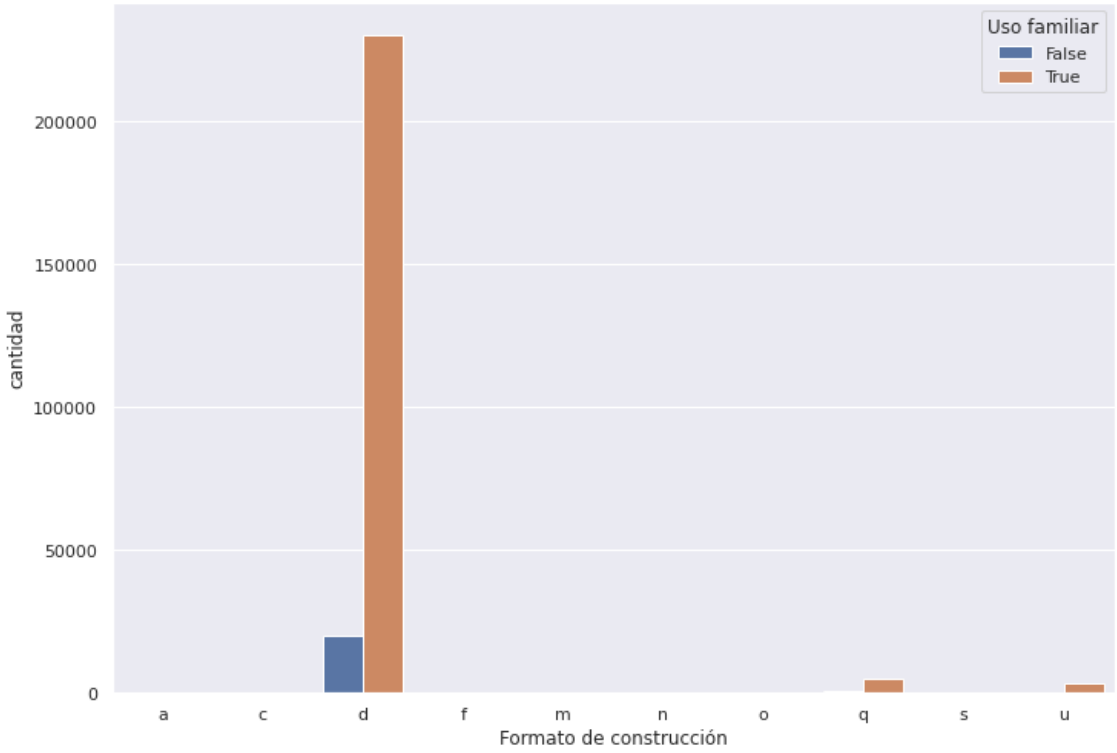
Cabe aclarar que, se intentó analizar conjuntamente los tres geo level con la cantidad de familias, pero fue una operación demasiado costosa para la herramienta utilizada (Google colab).

6.2.4. ¿Hay alguna relación entre el diseño sísmico y si la edificación se utilizó para uso familiar?

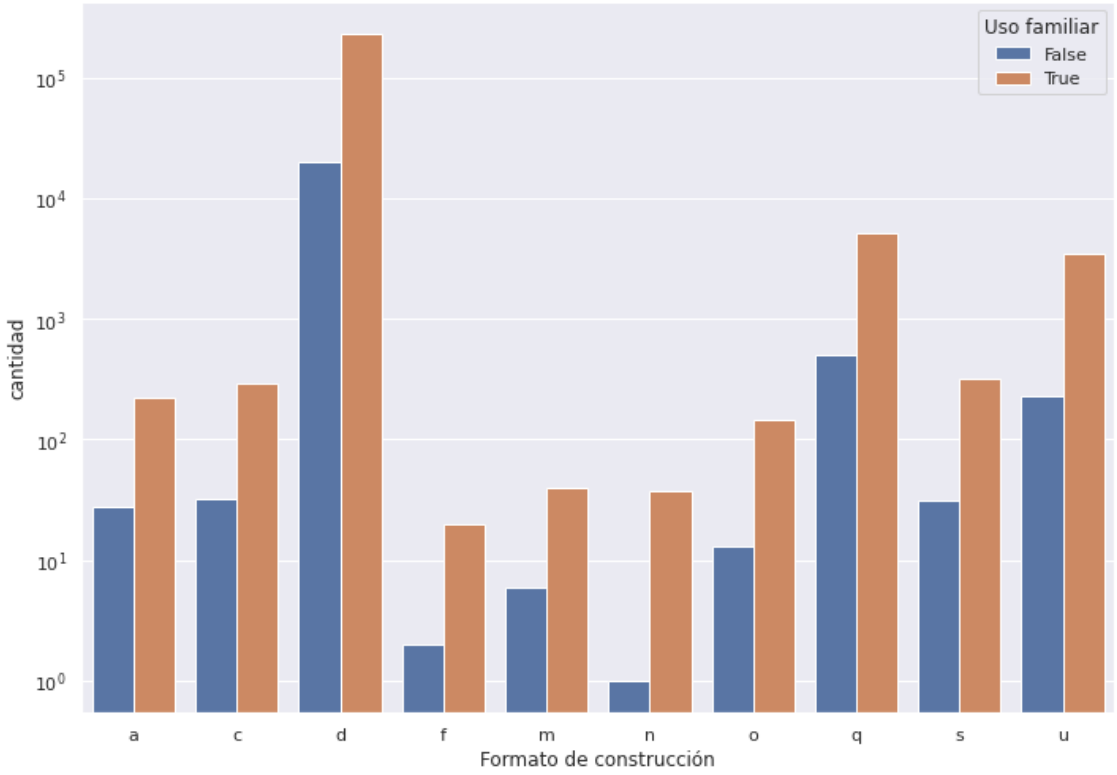
Se quiso analizar el formato de construcción de una edificación, y si era utilizada o no como vivienda familiar ($'count\_families' > 0$).

Vimos que el formato de tipo 'd' superaba con muchos órdenes de magnitudes al resto de los formatos. Por esta razón, también hicimos un gráfico en escala logarítmica.

Formato de construcción vs Cantidad de familias

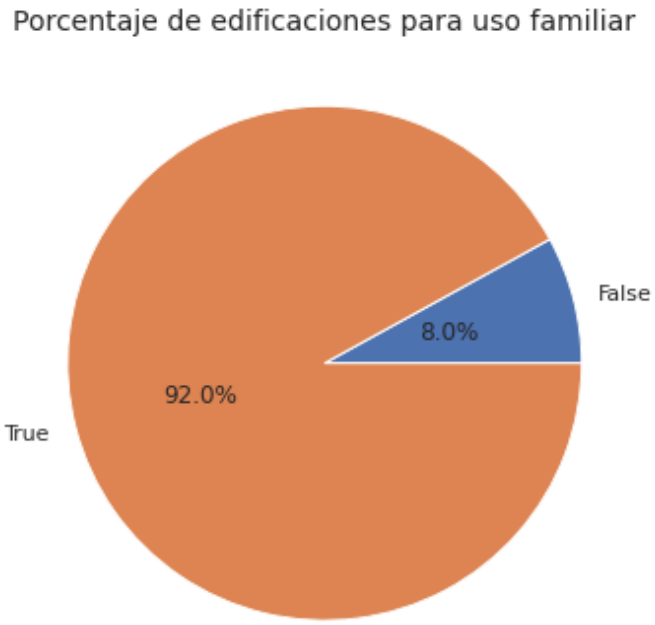


Formato de construcción vs Cantidad de familias - Escala logarítmica



6.2.5. ¿Cómo es la proporción de edificaciones de uso familiar con respecto a las que no?

Por último, para cerrar estos análisis, otra pregunta no relacionada con el daño, pero que también parece interesante conocer, es cuál es la proporción de edificios totales versus aquellos de uso familiar. Lo que vimos con los datos, es que únicamente el 8 % de las construcciones no eran de uso familiar.



6.3. Características de las edificaciones

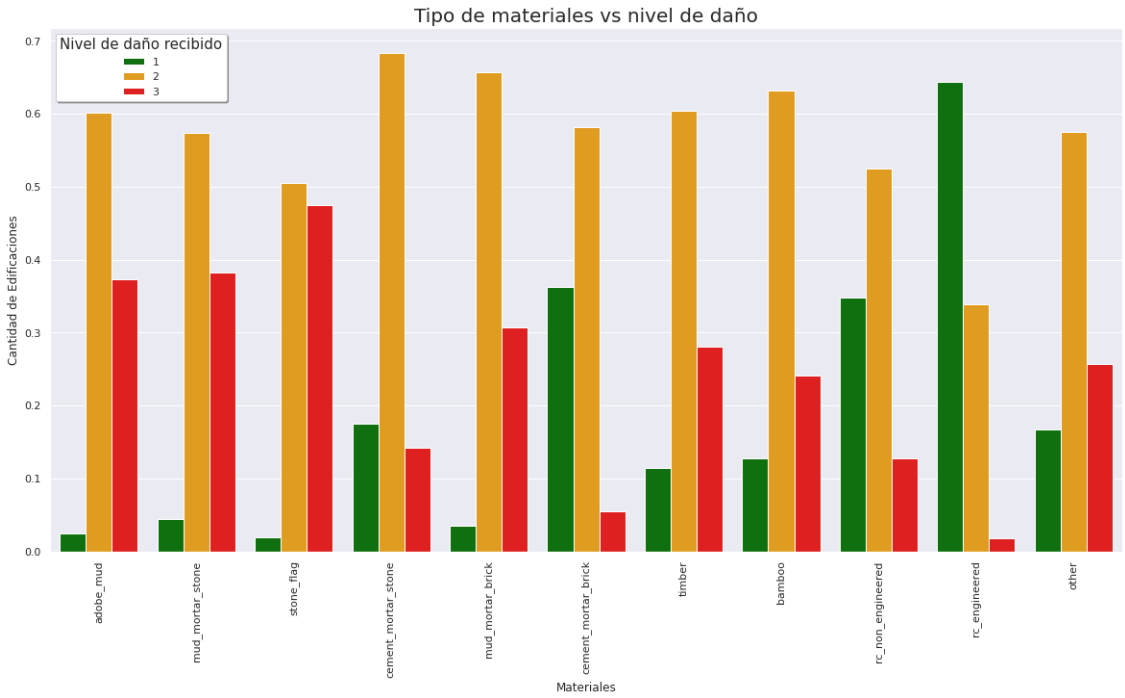
En esta sección estudiaremos como se relacionan las diferentes características de las edificaciones (antigüedad, materiales usados, etc) del país con el nivel de daño que recibieron por el terremoto. Para poder hacer el análisis de lo mencionado anteriormente, usamos las variables que describiremos a continuación:

- count_floors_pre_eq (tipo: entero): número de pisos en la edificación antes del terremoto.
- age (tipo: entero): antigüedad de la edificación en años.
- has_superstructure_adobe_mud (variable que indica si la edificación fue construida con adobe/barro.)
- has_superstructure_mud_mortar_stone (variable que indica si la edificación fue construida con barro - piedra.)
- has_superstructure_stone_flag (variable que indica si la edificación fue construida con piedra.)
- has_superstructure_cement_mortar_stone (variable que indica si la edificación fue construida con cemento - piedra.)
- has_superstructure_mud_mortar_brick (variable que indica si la edificación fue construida con barro - ladrillos.)

- has_superstructure_cement_mortar_brick (variable que indica si la edificación fue construida con cemento - ladrillos.)
- has_superstructure_timber (variable que indica si la edificación fue construida con Timber (madera específica para la construcción).)
- has_superstructure_bamboo (variable que indica si la edificación fue construida con Bambú (caña).)
- has_superstructure_rc_non_engineered (variable que indica si la edificación fue construida con concreto reforzado no-diseñado.)
- has_superstructure_rc_engineered (variable que indica si la edificación fue construida con concreto reforzado diseñado.)
- has_superstructure_other (variable que indica si la edificación fue construida con otro material.)
- Aclaración de las cantidades:
  Las cantidades mostradas son a nivel porcentual ya que decidimos normalizarlas.

6.3.1. ¿Hay algún material que haya sido mas resistente al terremoto?

A continuación podemos observar el gráfico realizado entre el tipo de material y el nivel de daño recibido.



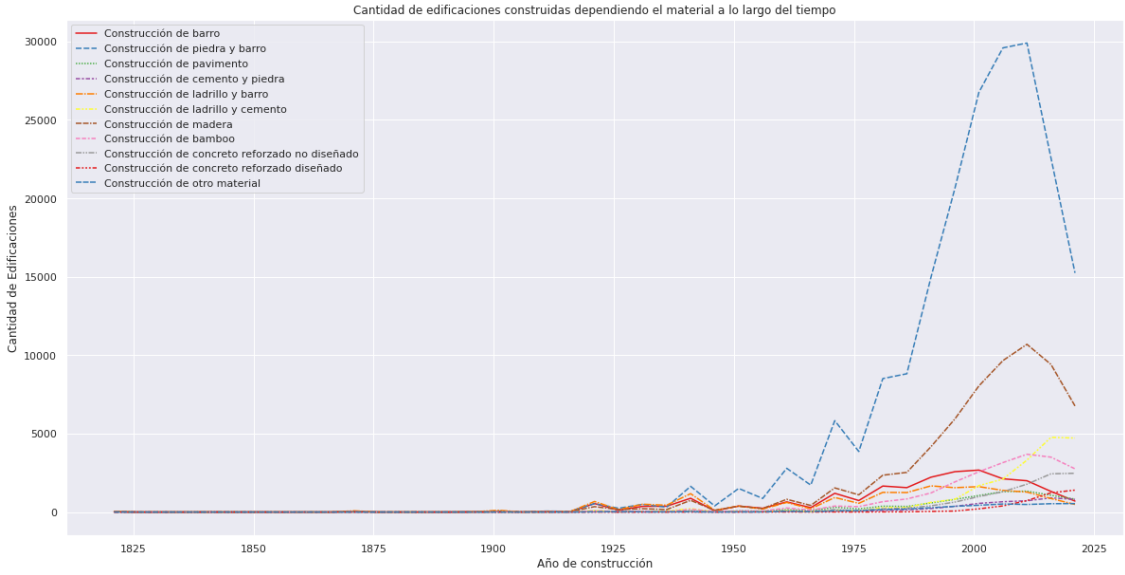
Analizando el gráfico se puede ver que en la mayoría de los materiales predomina un daño de nivel 2.
El material que parece ser mas resistente es el rc_engineered (concreto reforzado diseñado) ya que la mayoría del daño recibido en esas edificaciones construidas con este material es bajo (nivel 1) y en comparación muy pocas recibieron daño de nivel 3.

Nuestra conclusión es que este material (rc_engineered) fue el mas resistente al terremoto, ya

que el gráfico muestra que la mayoría de las edificaciones sufrieron daños del nivel 1 y nivel 2, y menos del 0.1 de las edificaciones sufrieron daños de nivel 3.

6.3.2. En caso de que haya un material mas resistente al terremoto, ¿Es el material qué predomina en las construcciones?

A continuación podemos observar un gráfico que muestra como fueron usados los materiales a lo largo del tiempo.



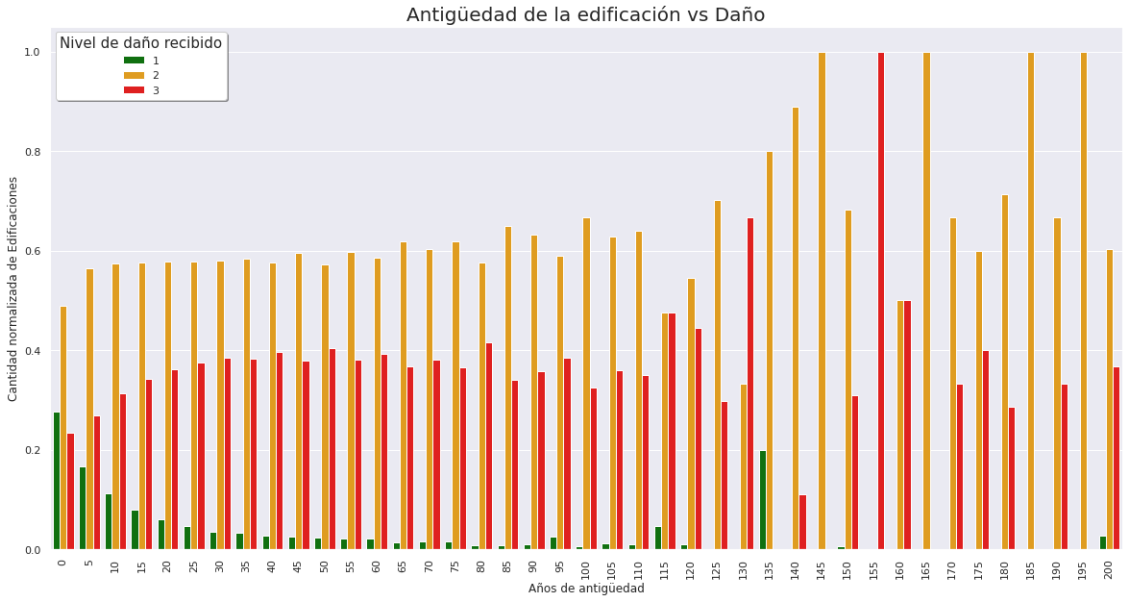
A partir del gráfico, podemos ver que predominan las construcciones de piedra y barro y en segundo lugar las construcciones con madera.

A continuación armamos una tabla con las cantidades exactas de construcciones con cada tipo de material desde el año 1825 hasta la actualidad con los valores ordenados de forma descendente.

Material	Cantidad de edificaciones
construcción de barro	197524
construcción de madera	65969
construcción de barro	22907
construcción de bamboo	22010
construcción de ladrillo y cemento	19534
construcción de ladrillo y barro	17629
construcción de concreto reforzado no diseñado	11054
construcción de pavimento	8879
construcción de cemento y piedra	4730
construcción de concreto reforzado diseñado	4113
construcción de otro material	3883

Podemos concluir que el material que mas resistió a los daños del terremoto no es el mas usado en las construcciones, de hecho es uno de los menos usados, ya que al observar la tabla lo encontramos en la ante-ultima posición.

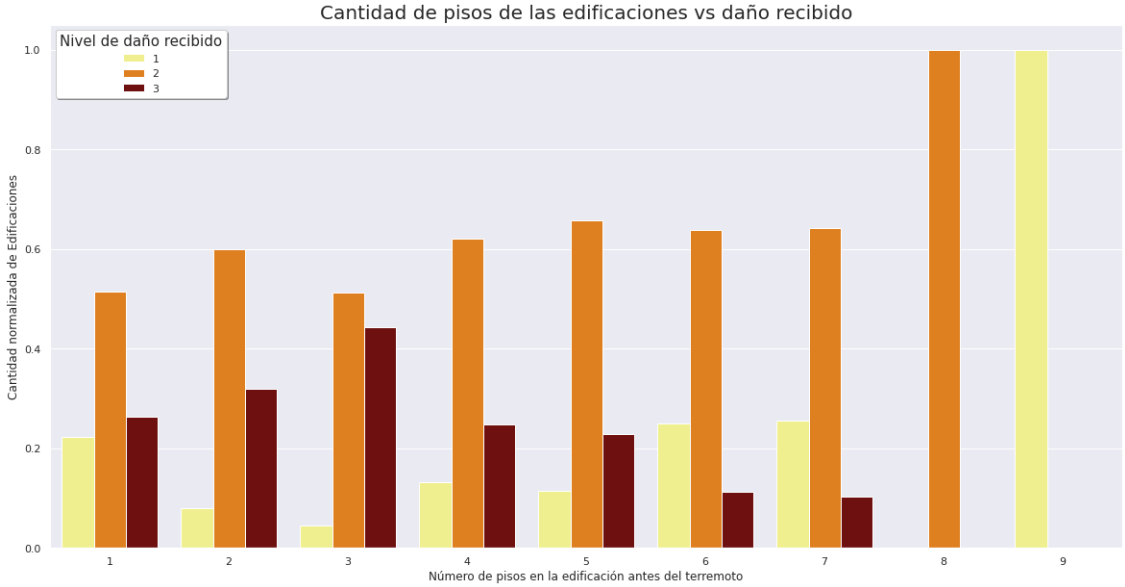
6.3.3. ¿La antigüedad de la edificación, influyó en el nivel de daño que recibió?



Por lo que se puede observar en el gráfico, la mayor cantidad de edificaciones sufrió mayor cantidad de daño de nivel 2 y 3 sin importar la antigüedad. También se puede ver cuanto mas nos acercamos a las edificaciones construidas actualmente hay menos daño de tipo 3 y este es reemplazado por daño de tipo 1. El daño de tipo 2 se mantiene relativamente constante a lo largo de la antigüedad del edificio.

6.3.4. Las edificaciones que tienen mayor cantidad de pisos, ¿sufrieron un daño menor que las que tienen mayor cantidad de pisos?

Como mencionamos anteriormente en la hipótesis, nosotros creemos que las edificaciones que cuentan con mas pisos serán las menos afectadas ya que creemos, son las que mejor preparadas están para los sismos.

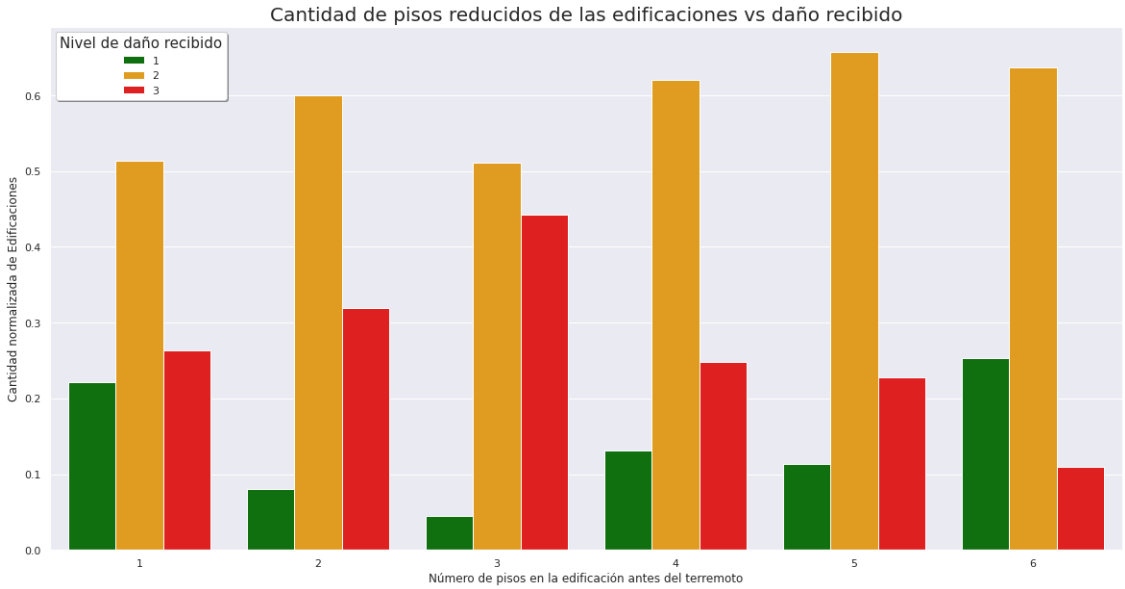


Como podemos observar en el gráfico, la cantidad de edificaciones con pisos 7, 8 y 9 es muy pequeña en comparación con el resto.

Adjuntamos una tabla para poder observar bien las cantidades.

Número de pisos	Cantidad de edificaciones
1	40272
2	155754
3	55332
4	5390
5	2218
6	204
7	39
8	1
9	1

La tabla demuestra claramente que las cantidades de edificaciones con pisos 6, 7, 8 y 9 son muy pocas y no nos da mucha información, por este motivo decidimos agrupar las edificaciones con cantidad de pisos mayores a 5 y procedimos a realizar otro gráfico de esta forma.



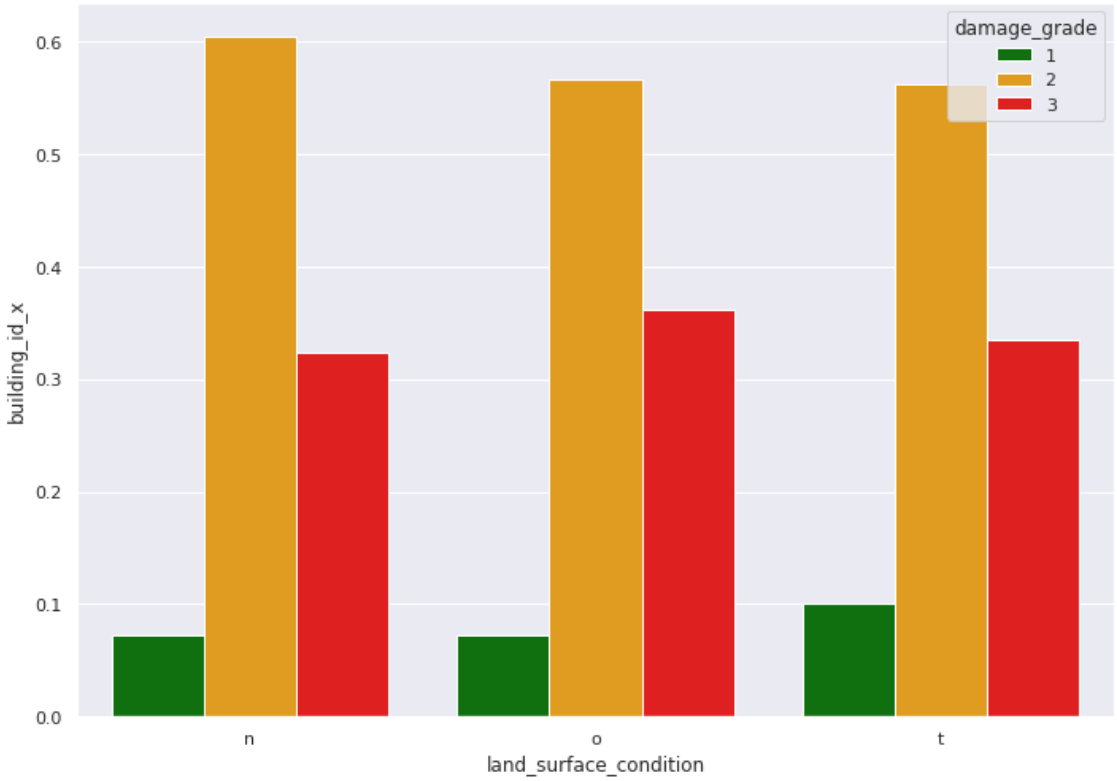
Con los pisos mayores a 5 agrupados, podemos ver que las construcciones con 3 pisos son las que mas daño sufrieron, ya que hubo mucho daño de nivel 2 y nivel 3.

Las edificaciones con pisos mayores a 5 son las que menos daño de tipo 3 sufrieron, y mas daño de tipo 1 y 2 sufrieron, por lo que nos hace concluir que aunque la cantidad de edificaciones con estos pisos sean pocas, son las que mejor preparadas están para soportar un sismo.

6.3.5. ¿Qué tipo de daño sufrieron las edificaciones dependiendo de la condición de la superficie en la que se encontraba?

Como mencionamos en el apartado de hipótesis, esperamos ver algún tipo de variación en los daños dependiendo de la condición de la misma.

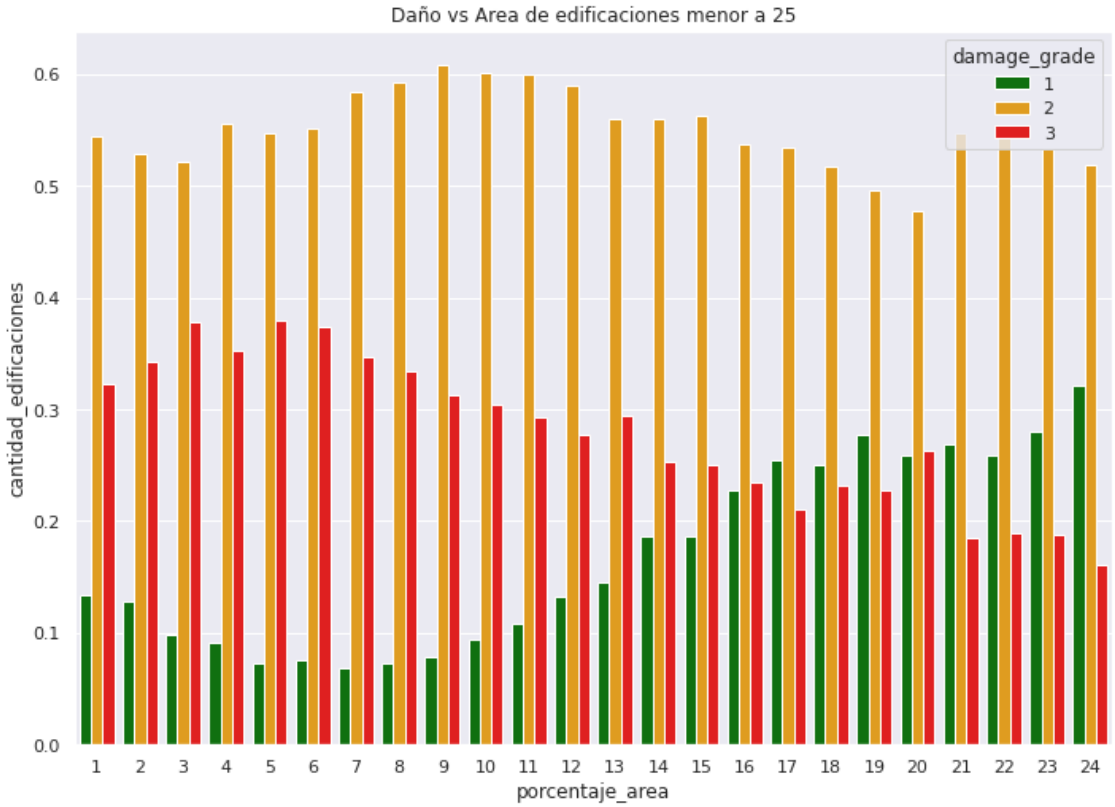
Daño vs Condición de la superficie

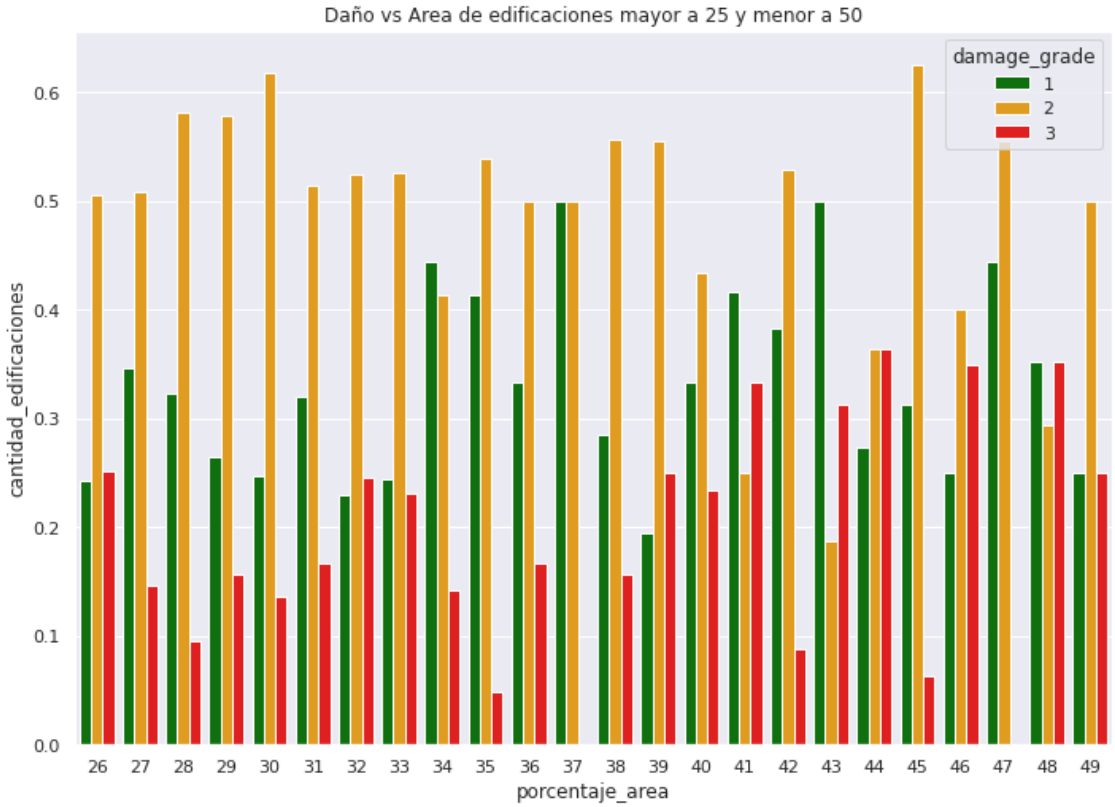


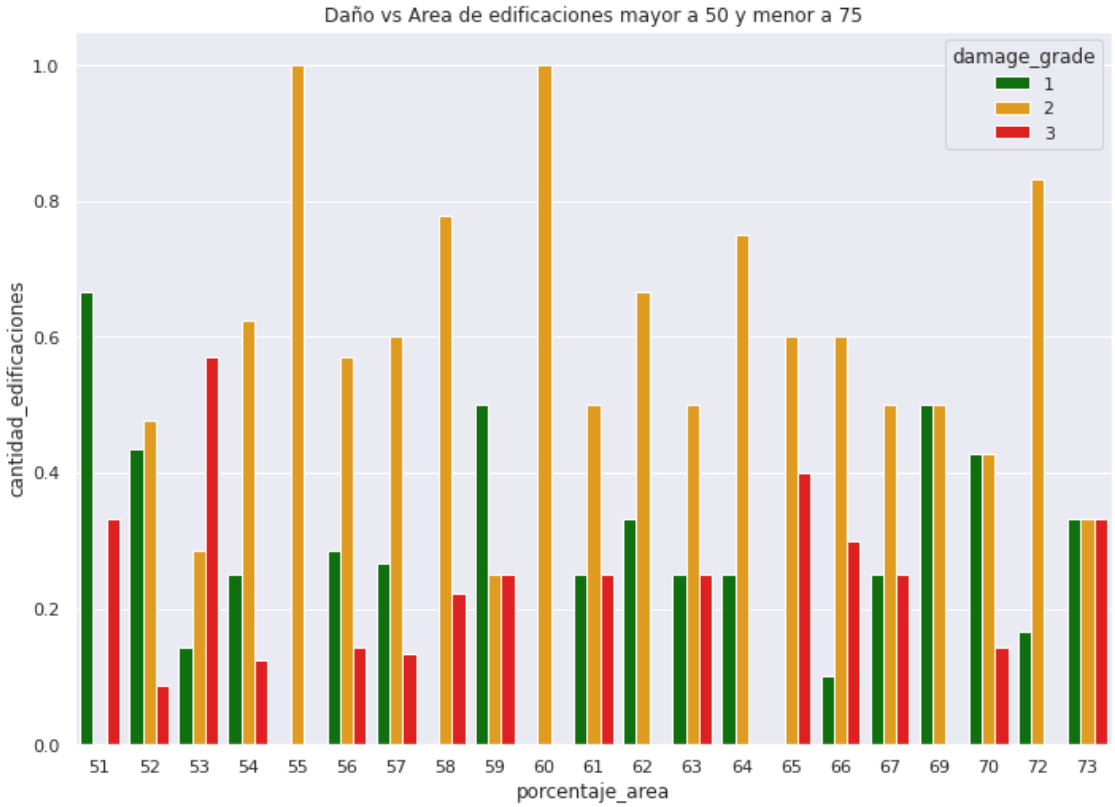
Se puede observar que la distribución del daño para las diferentes condiciones es similar, con lo cual es lo opuesto a lo que pensábamos.

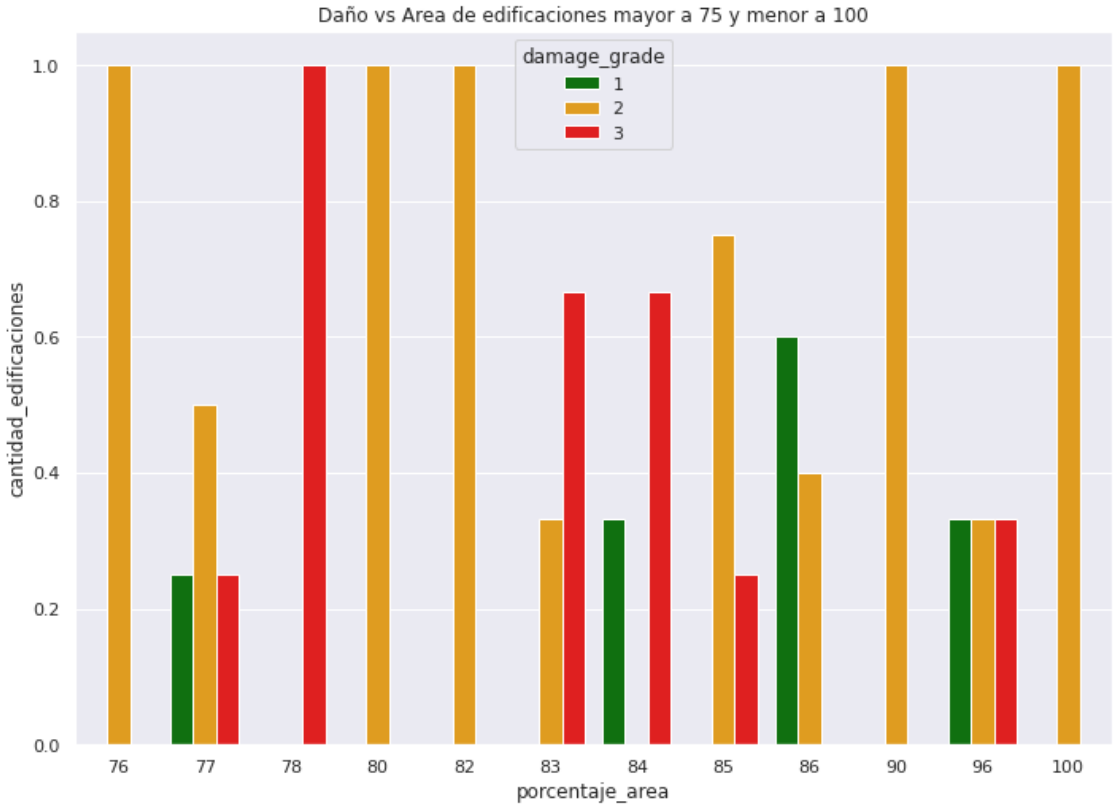
6.3.6. El área de la edificación influyó en el daño recibido?

En este apartado, dividimos el análisis en cuatro partes, ya que la variable area_percentage va de 0 a 100 y colocar todo en uno solo hace que la carga sea lenta y no se puedan visibilizar bien los datos.


Daño vs Area de edificaciones menor a 25
damage_grade
1
2
3
cantidad_edificaciones
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
porcentaje_area


Daño vs Area de edificaciones mayor a 25 y menor a 50
damage_grade
1
2
3
cantidad_edificaciones
0.0
0.1
0.2
0.3
0.4
0.5
0.6
26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
porcentaje_area


Daño vs Area de edificaciones mayor a 50 y menor a 75
damage_grade
1
2
3
cantidad_edificaciones
0.0
0.2
0.4
0.6
0.8
1.0
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 69 70 72 73
porcentaje_area


Daño vs Area de edificaciones mayor a 75 y menor a 100
damage_grade
1
2
3
cantidad_edificaciones
0.0
0.2
0.4
0.6
0.8
1.0
76 77 78 80 82 83 84 85 86 90 96 100
porcentaje_area

Se consideró filtrar por aquellas edificaciones que tienen, por lo menos, 10 datos para cada valor de área.

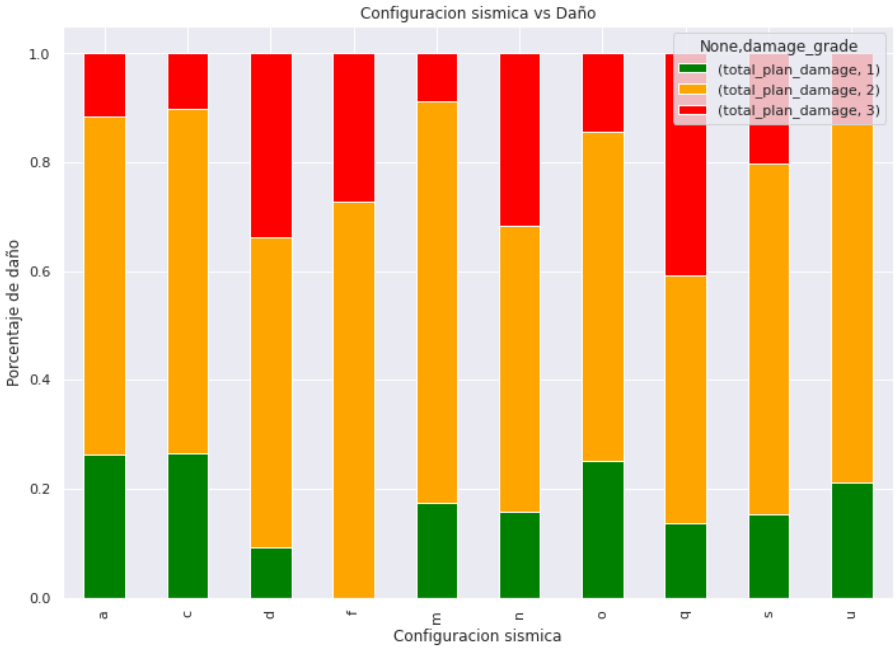
Podemos observar que el daño 3 crece en los primeros valores del área y luego comienza a decrecer, manteniéndose por debajo del daño que sufre inicialmente. El daño de tipo 2 se mantiene alto para cualquier valor, pero a medida que avanza se producen altibajos, esto puede deberse a que la muestra en ese valor puede ser baja.

6.4. Diseño sísmico

En este apartado estudiaremos las relaciones que tienen las siguientes variables con el tipo de diseño sísmico que tenia la edificación:

- damage_grade (daño que sufrió la edificación, puede ser 1 (bajo), 2 (medio), 3 (alto).
- has_secondary_use_* (se analizaron todas las variables con este prefijo, representa el tipo de uso que se le daba a la vivienda).
- land_surface_condition (condición de la superficie terrestre donde fue construida la edificación)

6.4.1. ¿Qué tipo de daño sufrieron las edificaciones dependiendo de la configuración sísmica adoptada?



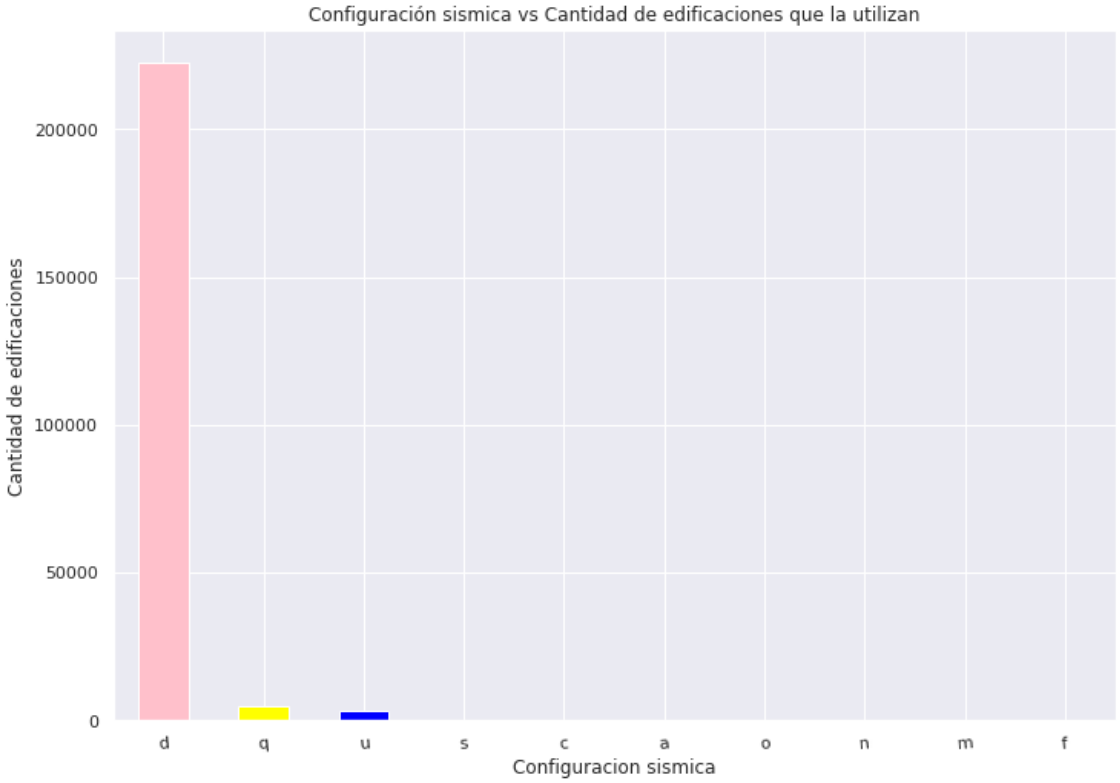
En principio no hay ninguna configuración que haya soportado totalmente los daños más graves, tampoco podemos ver alguna en la que predomine mas el daño 1 que 2 ó 3. Si podemos observar que las que tenían configuraciones 'm','c' y 'a' tuvieron, en menor proporción, daños de tipo 3.

En cuanto a las más afectadas, destacan, 'f' que no tuvo ningún tipo de daño 1, 'q' que es la única que el daño 3 casi supera al daño 2. En 'n' y 'd' también se ve una proporción grande de daño 3.

6.4.2. ¿Qué tipo de daño sufrieron las edificaciones de uso familiar dependiendo de la configuración sísmica adoptada?

Queremos ver si hay alguna configuración que predomine para las casas en las que viven familias, creemos que se puede observar una tendencia para una configuración, tal vez porque el material es más barato o el único al que pueden acceder.

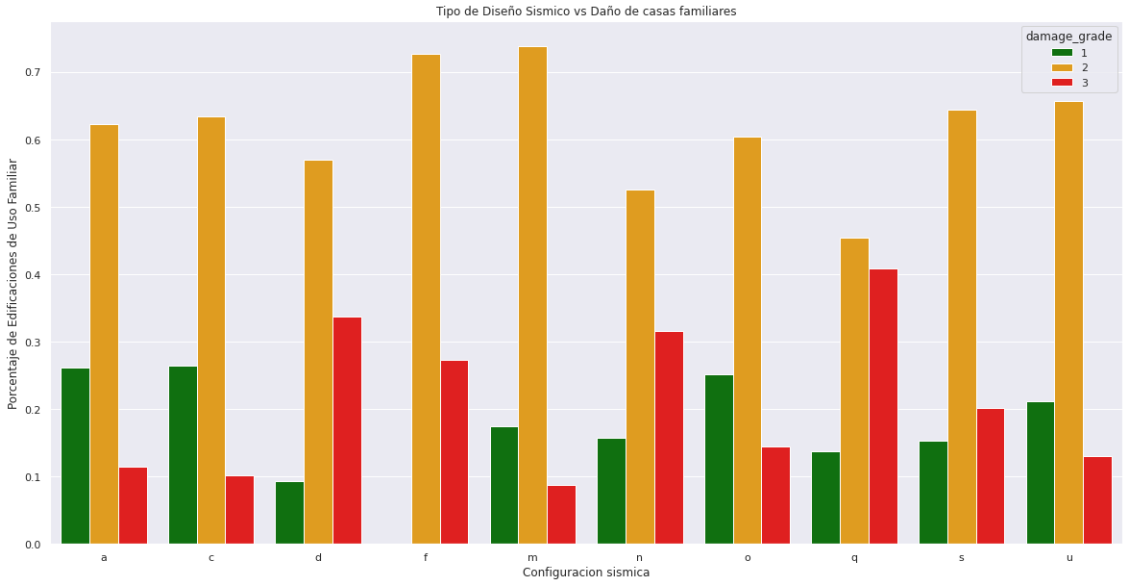
A continuación mostraremos un gráfico de cual es la configuración mas utilizada por las familias.



En cantidades:

Configuración	Cantidad
d	222352
q	5064
u	3089
s	296
c	221
a	208
o	129
n	34
m	32
f	20

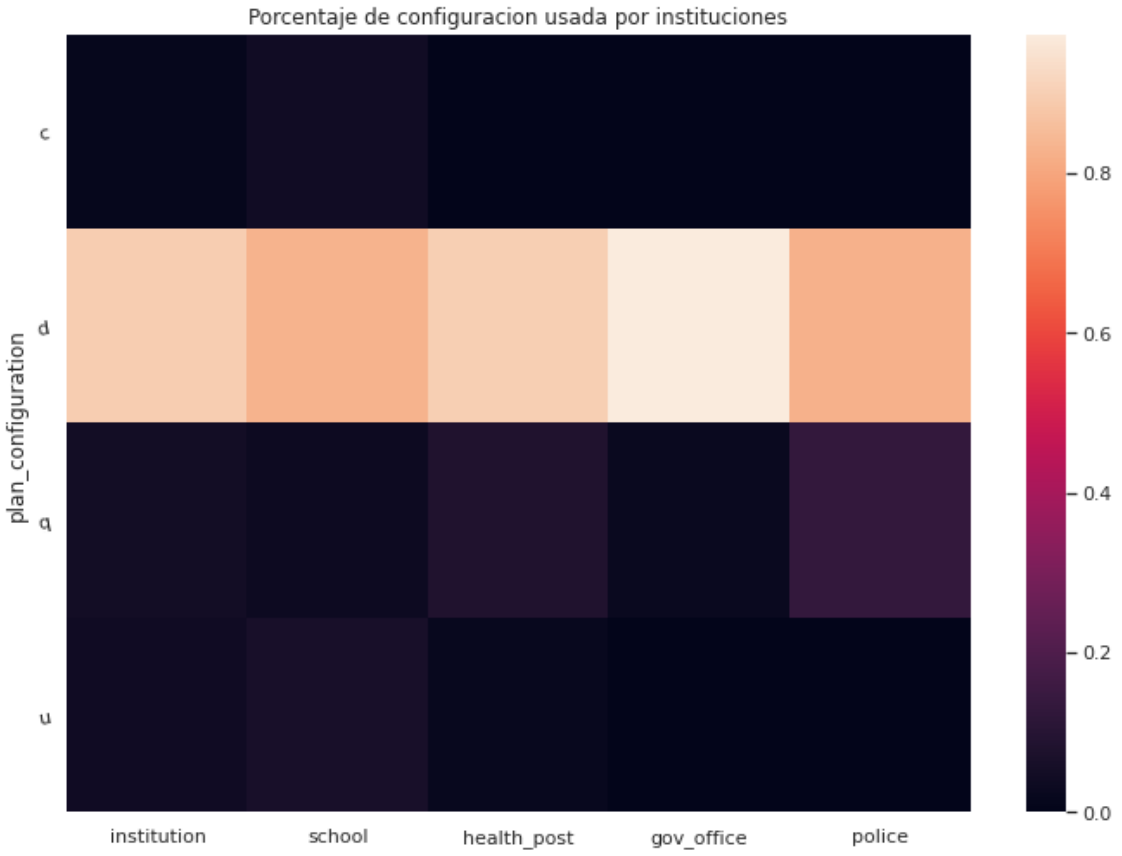
Como creíamos, hay una que predomina, y es la 'd'.



Con las cantidades observadas en el primer gráfico, la 'd', que es la mas utilizada es una de las que peores consecuencias tuvo, junto a 'q', que es la segunda más utilizada. Por otro lado, 'u' que es la tercera, si bien sufrió daños graves, es menor a las dos anteriores. No hay una configuración que nos haga pensar que se podrían evitar daños medios o graves al utilizarla.

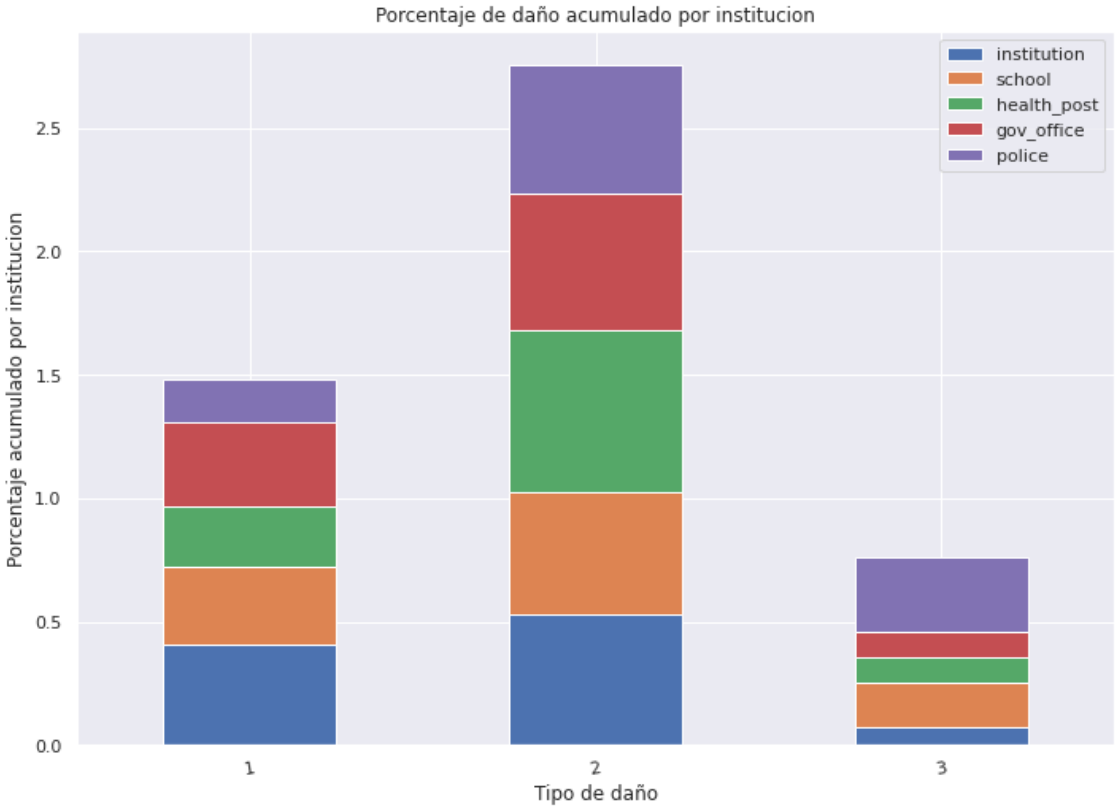
6.4.3. ¿Existirá alguna Institución que utilice alguna configuración en particular?

Suponemos que al ser una Institución podría tener mejor diseño ya que el capital para su construcción es probable que haya sido mayor.



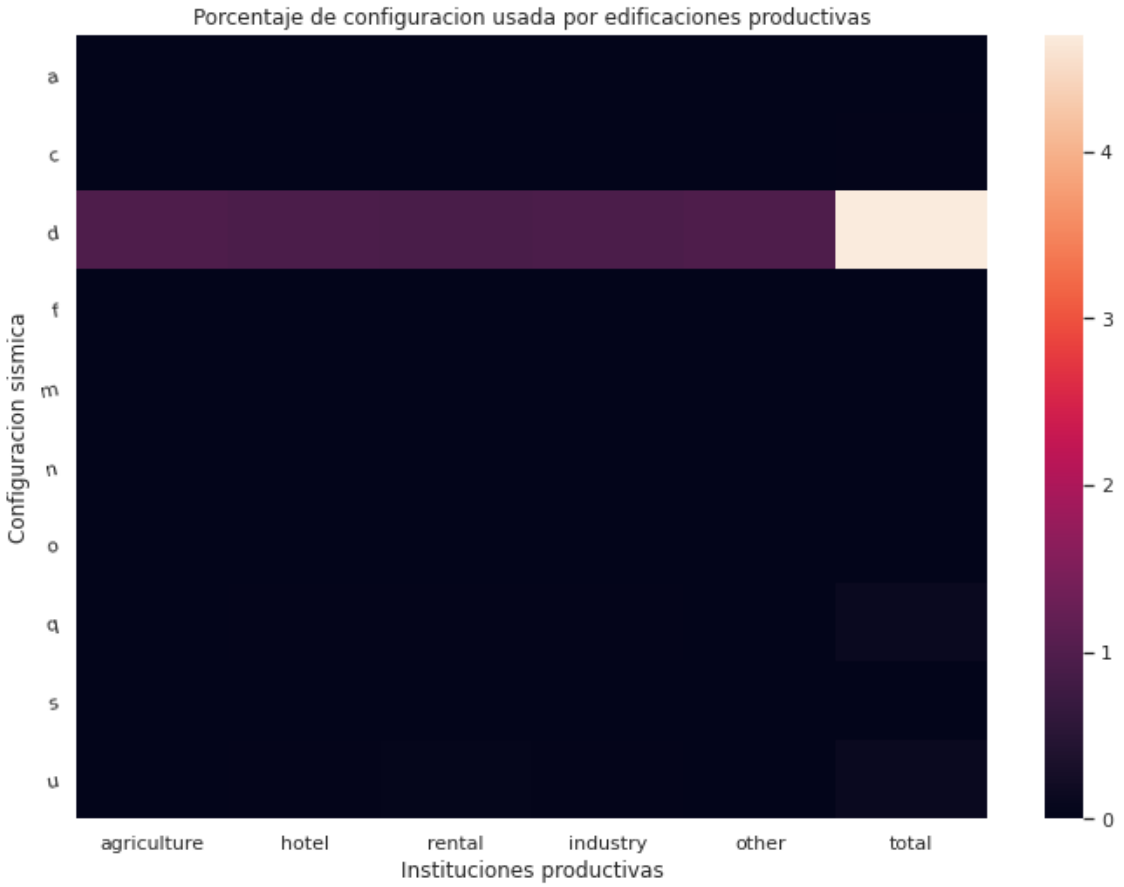
No hay diferencia con las de uso familiar, institution, school, health_post, gov_office y police están plenamente relacionadas con la configuración 'd'. Tal vez las estaciones de policía, aproximadamente el 13 % usa 'q'.

Veamos el daño que sufrieron las instituciones



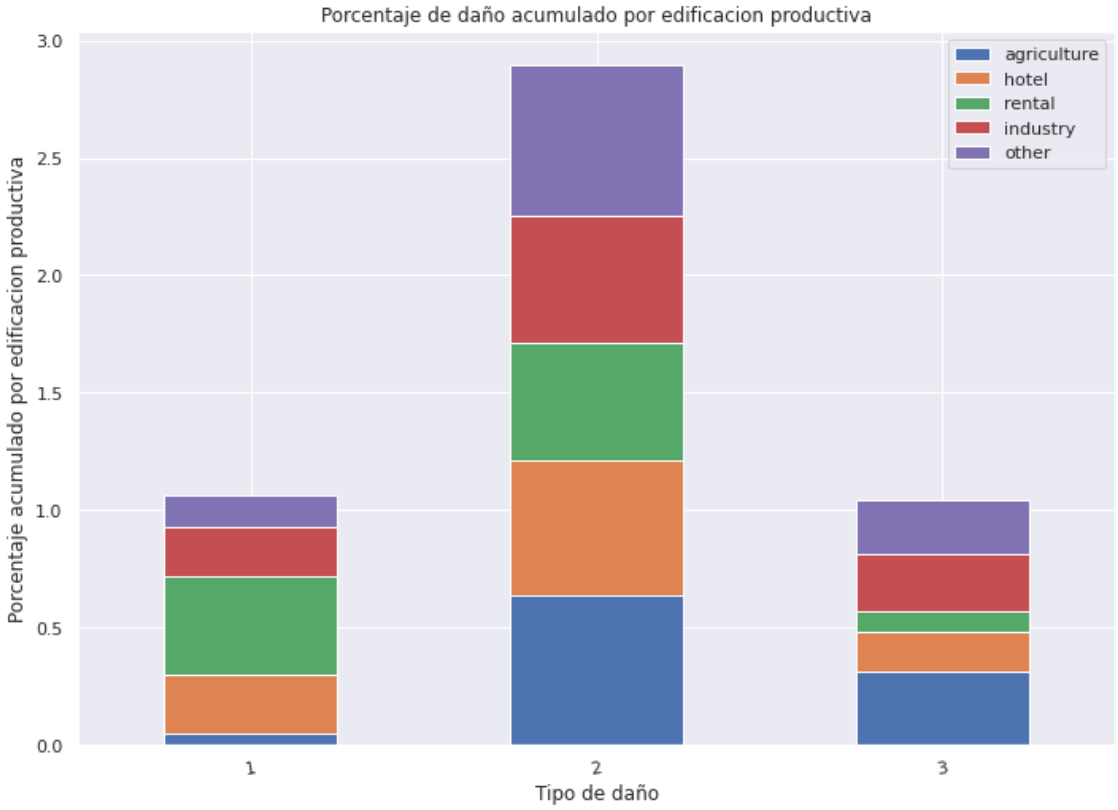
6.4.4. ¿Variará la configuración en las edificaciones de uso productivo?

Ya vimos que en las instituciones no estaban distribuidas las configuraciones, esperamos ver que con los edificios de uso productivo, al menos en industrias y hoteles, no se usa únicamente la configuración 'd'. Veamos..



Sorpresa, otra vez predomina 'd'. Intentaremos analizar mas adelante por qué se utiliza en más del 96 % de las edificaciones.

Observemos sus daños



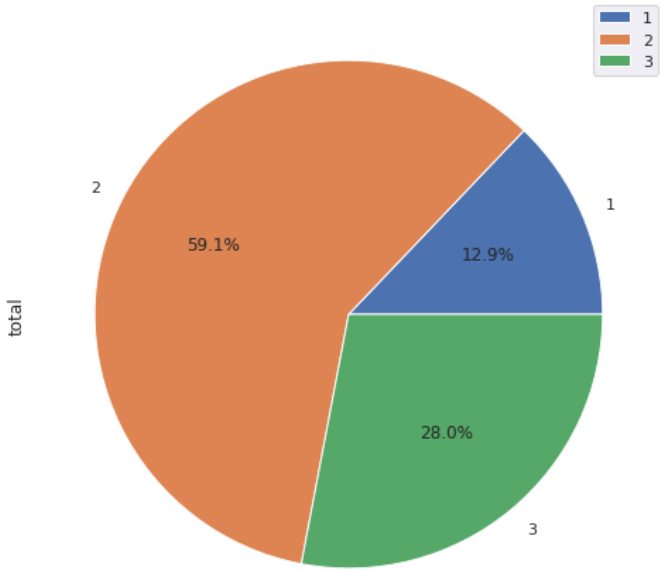
Otra vez, al contrario de lo que pensábamos, los hoteles y industrias sufrieron bastante daño. En cambio rental, fue la que tuvo menor daño 2 y 3.

7. Análisis de edificaciones con edad 995

7.1. Análisis del daño

Las edificaciones que tienen este valor en el dataset representan, aproximadamente, un 0.05 % del total.

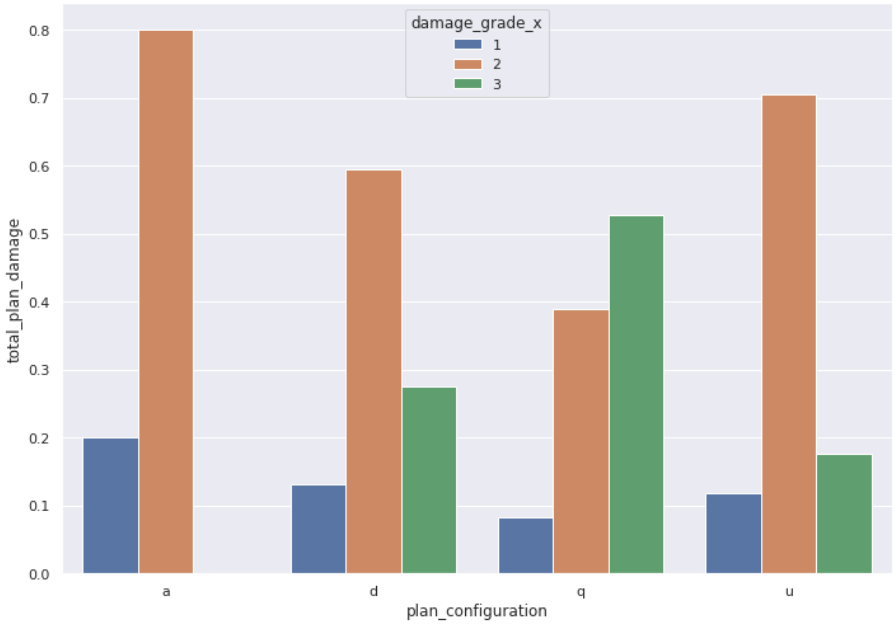
El porcentaje de tipo de daño que sufrieron se ve a continuación:



Se puede observar que hubo mas casos de daño 2 y mas de 3 que de 1.

Ahora queremos ver si estas edificaciones con un valor 'extraño' presentan algún tipo de configuración sísmica que haga que sean propensas a sufrir más o menos daño.

7.1.1. ¿Qué configuración sísmica tenían estos edificios?

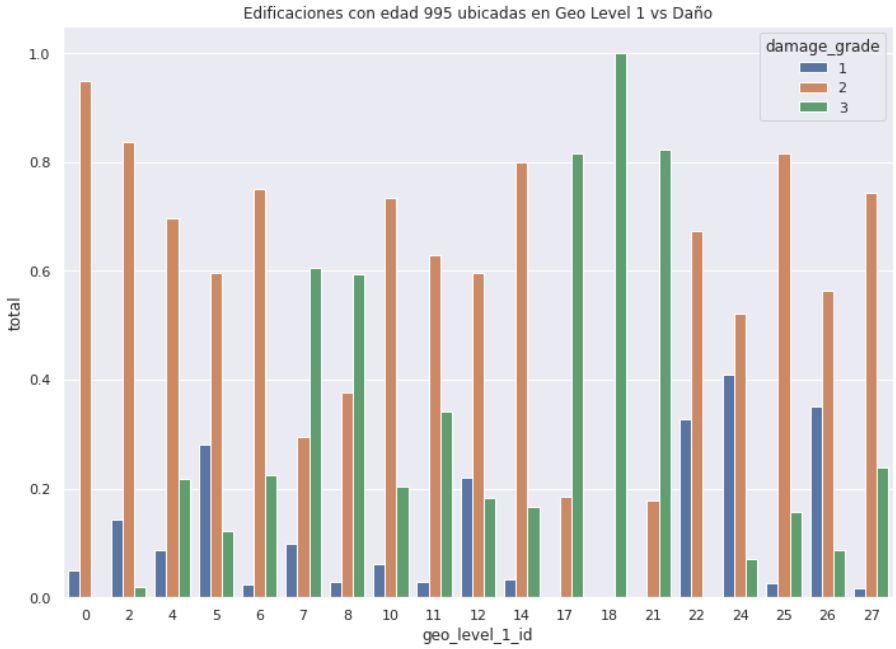


A simple vista se ve que la configuración 'a' no tiene absolutamente ningún daño 3. Nos preguntamos si será por que, tal vez, no hay una buena cantidad que permita analizar la variable?

Bien, al obtener la cantidad de edificaciones para esta variable, resulta que hay una muestra de tan solo 5, lo cual resulta muy pequeña como para poder sacar conclusiones precisas. Aunque a priori, no es propensa a daños graves.

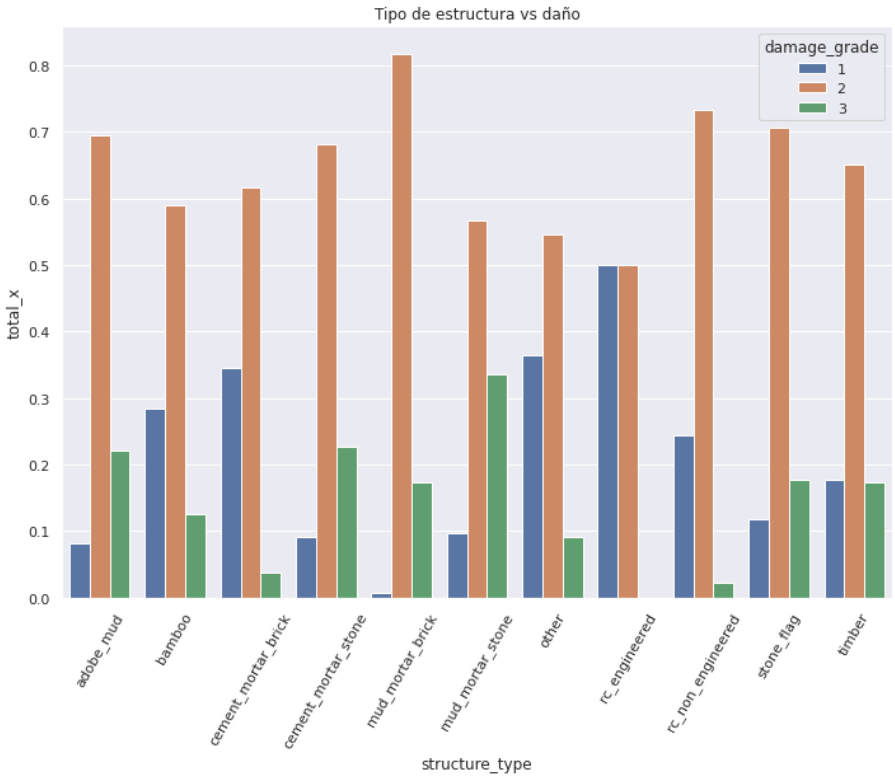
Por otro lado, la configuración 'q', con una muestra de 36, destaca por los demás daños, el 3.

7.1.2. ¿En que región de geo level 1 estaban ubicadas y que tipo de daño sufrieron?



Vemos que las que se encontraban en las regiones 7,8 17,18, 21 sufrieron mayoritariamente daños de tipo 3, lo cual tiene sentido, ya que coinciden con las zonas analizadas al inicio del trabajo.

7.1.3. ¿Con qué material fueron construidas estas edificaciones?



8. Conclusiones adicionales

Adicionalmente, a las conclusiones hechas en cada sección sumamos las siguientes:

- Podemos concluir que las zonas con Geo level 1 ID 17, 18 y 21 fueron las mas afectadas y hay mas posibilidades que una edificación haya sido destruida completamente por el terremoto. También podemos decir que no hubo ninguna zona donde predomine el daño de nivel 1.
- Como mas del 95 % de las construcciones tienen formato de diseño sísmico "d", esta puede ser la razón de porque el daño es tan parejo a lo largo de todos los valores de geo level 1.